Adaptive variable selection in nonparametric sparse additive models

Cristina Butucea[1,2] and Natalia Stepanova[3]

[1] Université Paris-Est Marne-la-Vallée,
LAMA(UMR 8050), UPEM, UPEC, CNRS, F-77454, Marne-la-Vallée, France

[2] ENSAE-CREST-GENES 3, ave. P. Larousse 92245 MALAKOFF Cedex, FRANCE

[3] School of Mathematics and Statistics, Carleton University, Ottawa, Ontario, K1S 5B6 Canada

Abstract

We consider the problem of recovery of an unknown multivariate signal f observed in a d-dimensional Gaussian white noise model of intensity ε. We assume that f belongs to a class of smooth functions $\mathcal{F}^d \subset L_2([0,1]^d)$ and has an additive sparse structure determined by the parameter s, the number of non-zero univariate components contributing to f. We are interested in the case when $d = d_\varepsilon \to \infty$ as $\varepsilon \to 0$ and the parameter s stays "small" relative to d. With these assumptions, the recovery problem in hand becomes that of determining which sparse additive components are non-zero. Attempting to reconstruct most non-zero components of f, but not all of them, we arrive at the problem of almost full variable selection in high-dimensional regression. For two different choices of $\mathcal{F}^d$, we establish conditions under which almost full variable selection is possible, and provide a procedure that gives almost full variable selection. The procedure does the best (in the asymptotically minimax sense) in selecting most non-zero components of f. Moreover, it is adaptive in the parameter s.

Keywords and phrases: high-dimensional nonparametric regression; sparse additive signals; adaptive variable selection; exact and almost full selectors

Mathematics Subject Classification: 62G08, 62G20

1 Introduction

In recent years, there has been much work on methods for variable selection in high dimensional settings; refer, for example, to [3, 6, 8, 18] and references therein. Among a variety of methods proposed, the lasso has become an important tool for sparse high-dimensional regression problems. Motivated by the fact that finding the lasso solutions is computationally demanding, Genovese et al. [6] studied the relative statistical performance of the lasso and marginal regression, which is also known as simple thresholding, for sparse high-dimensional regression problems. They found that marginal regression, where each dependent variable is regressed separately on each covariate, provides a good alternative to the lasso, and concluded that their procedure merits further study. Handling the problem of reconstruction in high dimensional regression, Genovese et al. [6] distinguished between the cases of exact, almost full, and no recovery. Exact recovery refers to the situation where the set of all relevant components can be consistently recovered (asymptotically). Almost full recovery stands for the possibility of having the number of misclassified components negligibly small as compared to the number of all relevant components. The latter strategy requires milder restrictions on a statistical model and can be used in the situations where exact recovery is impossible. If neither exact nor almost full recovery can be achieved, we speak of 'no recovery' when the optimal risk is as large as the number of relevant components and any recovery procedure fails completely.

Ingster and Stepanova [13] extended the idea of Genovese et al. [6] to the case of nonparametric regression. Specifically, they addressed the problem of recovering sparse additive smooth signals observed in the continuous regression model and showed that, asymptotically, as dimension increases indefinitely, exact variable selection is possible and is provided by a suitable thresholding procedure. The procedure in [13] is optimal in the asymptotically minimax sense. It is also free from the sparsity parameter and thus is adaptive. At the same time, the more intricate problem of almost full recovery in an adaptive setup remained unsolved. We shall treat this problem in the present paper.

Our setting is that of a multivariate signal $f \in \mathcal{F}^d \subset L_2([0,1]^d) = L_2^d$ corrupted by a Gaussian white noise of a given intensity ε:

$$X_\varepsilon = f + \varepsilon W, \tag{1}$$

where W is a d-dimensional Gaussian white noise on $[0,1]^d$, $\varepsilon > 0$ is a noise intensity, and $\mathcal{F}^d$ is a subset of L_2^d that consists of sufficiently smooth functions. In the present paper, two examples of $\mathcal{F}^d$ will be considered. In this model, the "observation" is the function $X_\varepsilon \colon L_2^d \to \mathcal{G}$ taking its values in the set $\mathcal{G}$ of normal random variables such that if $\xi = X_\varepsilon(\phi)$, $\eta = X_\varepsilon(\psi)$, where $\phi, \psi \in L_2^d$, then $\mathbf{E}(\xi) = (f, \phi)$, $\mathbf{E}(\eta) = (f, \psi)$, and $\mathbf{Cov}(\xi, \eta) = \varepsilon^2(\phi, \psi)$. For any $f \in L_2^d$, the observation X_ε determines the Gaussian measure $\mathbf{P}_{\varepsilon,f}$ on the Hilbert space L_2^d with mean function f and covariance operator $\varepsilon^2 I$, where I is the identity operator (see [9, 19] for references). The expectation that corresponds to the probability measure $\mathbf{P}_{\varepsilon,f}$ is denoted by $\mathbf{E}_{\varepsilon,f}$. In this paper, the case of growing dimension $d = d_\varepsilon \to \infty$ as $\varepsilon \to 0$ is studied. It is well known that the continuous model (1) serves as a good approximation to a more realistic equidistant sampling scheme with discrete Gaussian white noise. In such an approximation, ε^{-2} roughly corresponds to the number n of observations per unit cube $[0,1]^d$.

An important problem in this context is to recover f from noisy data. Attempting to suppress the curse of dimensionality and complement the findings in [13], we assume that f has an additive sparse structure. Our goal is to study under what conditions and by means of what procedure *almost full* recovery of an additive sparse signal f is possible. In other words, we wish to correctly identify most non-zero components of f. In doing so, we aim at providing the procedure that, for the two function spaces $\mathcal{F}^d$ of our interest, one consisting of functions of finite smoothness and the other consisting of functions of infinite smoothness, is optimal in the asymptotically minimax sense. In the almost full recovery regime, one can detect even smaller relevant components but, unfortunately, at the price of a loss in the rate. Therefore constructing the corresponding procedure is technically more demanding as compared to that in the exact recovery case. To develop a good almost full recovery procedure, we will use results from minimax hypothesis testing and minimax estimation theory.

To fix some notation and assumptions, let the signal f in model (1) be of the form (see, for example, [5] and [13])

$$f(\mathbf{x}) = \sum_{j=1}^d \eta_j f_j(x_j), \quad \mathbf{x} = (x_1, \dots, x_d) \in [0,1]^d, \quad \eta = (\eta_1, \dots, \eta_d) \in \mathcal{H}_{d,s},$$

where for a number $s \in \{1, \ldots, d\}$, called the *sparsity parameter*,

$$\mathcal{H}_{d,s} = \{\eta = (\eta_1, \ldots, \eta_d) : \eta_j \in \{0,1\}, 1 \leq j \leq d, \sum_{j=1}^{d} \eta_j = s\}.$$

The η_j's are non-random quantities taking values 0 and 1; the case $\eta_j = 1$ ($\eta_j = 0$) corresponds to the situation when the component f_j is active (non-active). When $s = o(d)$ we speak of a *sparse* additive signal f. In addition, each component f_j is assumed to be an element of a certain smooth function space $\mathcal{F}_\sigma \subset L_2[0,1]$ depending on a *known* parameter $\sigma > 0$; two examples of $\mathcal{F}_\sigma$ under study are introduced in Section 2. Thus, the class of s-sparse multivariate signals of interest is

$$\mathcal{F}^d_{s,\sigma} = \left\{ f : f(\mathbf{x}) = \sum_{j=1}^{d} \eta_j f_j(x_j),\ \int_0^1 f_j(x)\, dx = 0,\ f_j \in \mathcal{F}_\sigma,\ 1 \leq j \leq d,\ \eta = (\eta_j) \in \mathcal{H}_{d,s} \right\},$$

where the components satisfy side condition that guarantees uniqueness, and the signal recovery problem becomes that of determining which sparse additive components are non-zero.

In the context of variable selection, the problem of reconstruction of an additive function f is now stated as follows. For each component f_j of a signal $f \in \mathcal{F}^d_{s,\sigma}$, consider testing the hypothesis of no signal $H_{0j} : f_j = 0$ versus the alternative $H_{1j} : f_j \in \mathcal{F}_\sigma(r_\varepsilon)$, where for a positive family $r_\varepsilon \to 0$

$$\mathcal{F}_\sigma(r_\varepsilon) = \{g \in \mathcal{F}_\sigma : \|g\|_\sigma \leq 1,\ \|g\|_2 \geq r_\varepsilon\}, \tag{2}$$

and $\|\cdot\|_\sigma$ is a norm on $\mathcal{F}_\sigma$. In this problem, a precise demarcation between the signals that can be detected with error probabilities tending to 0 and the signals that cannot be detected is given in terms of a *detection boundary*, or *separation rate*, $r^*_\varepsilon \to 0$ as $\varepsilon \to 0$. For various function classes frequently used in minimax hypothesis testing, sharp asymptotics for r^*_ε are available (see, for example, [10]). The hypotheses H_{0j} and H_{1j} *separate asymptotically* (that is, the minimax error probability tends to zero) if $r_\varepsilon / r^*_\varepsilon \to \infty$ as $\varepsilon \to 0$. The hypotheses H_{0j} and H_{1j} *merge asymptotically* (that is, the minimax error probability tends to one) if $r_\varepsilon / r^*_\varepsilon \to 0$ as $\varepsilon \to 0$.

When H_{0j} and H_{1j} separate asymptotically, we say that f_j is *detectable.* If the hypotheses H_{0j} and H_{1j} separate (merge) asymptotically when $\liminf r_\varepsilon / r^*_\varepsilon > 1$ ($\limsup r_\varepsilon / r^*_\varepsilon < 1$), the detection boundary r^*_ε is said to be *sharp.* The knowledge of a sharp detection boundary r^*_ε allows us to have a meaningful problem of testing $H_{0j} : f_j = 0$ versus $H_{1j} : f_j \in \mathcal{F}_\sigma(r_\varepsilon)$ by choosing r_ε so that $\liminf_{\varepsilon \to 0} r_\varepsilon / r^*_\varepsilon > 1$. Otherwise, the function f_j will be too "small" to be noticeable.

Let us agree to say that any measurable function $\eta^* = \eta^*(X_\varepsilon)$ taking values on $\{0,1\}^d$ is a *selector.* Following [6] and [13], we judge the quality of a selector η^* of vector $\eta \in \mathcal{H}_{d,s}$ by using the *Hamming distance* on $\{0,1\}^d$, which counts the number of positions at which $\eta^* = (\eta^*_1, \ldots, \eta^*_d)$ and $\eta = (\eta_1, \ldots, \eta_d)$ differ:

$$|\eta^* - \eta| = \sum_{j=1}^{d} |\eta^*_j - \eta_j|.$$

Following [6], we distinguishe between exact and almost full recovery. Roughly, a selector $\eta^* = \eta^*(X_\varepsilon)$ is asymptotically *exact* if its maximum risk is $o(1)$. Likewise, a selector $\eta^* = \eta^*(X_\varepsilon)$

is asymptotically *almost full* if its maximum risk is $o(s)$ with s being the number of non-zero components f_j of a signal $f = \sum_{j=1}^{d} \eta_j f_j$.

Ingster and Stepanova [13] have obtained adaptive procedure that gives asymptotically exact reconstruction of a σ-smooth signal $f \in \mathcal{F}^d_{s,\sigma}$ observed in a d-dimensional Gaussian white noise model. A similar result for the space of infinitely-smooth functions is stated in this paper in Section 4.2 (see Theorems 1 and 2). Although the selector in Section 4.2 is based on somewhat different statistics when compared to the one in [13], both selectors have one common feature that their thresholds are free of the sparsity parameter s and therefore automatically adapt themselves to its values.

The goal of this paper is three-fold. First, we find a sharp detection boundary that allows us to separate detectable components of a signal $f \in \mathcal{F}^d_{s,\sigma}$ from non-detectable ones. Next, assuming that all active components f_j are detectable and that s belongs to a set $\mathcal{S}_d$, which puts some mild restrictions on the range of s, we construct a selector $\eta^* = \eta^*(X_\varepsilon)$ with the property

$$\sup_{s\in\mathcal{S}_d}\sup_{\eta\in\mathcal{H}_{d,s}}\sup_{f\in\mathcal{F}^d_{s,\sigma}} s^{-1}\mathbf{E}_{f,\eta}|\eta^* - \eta| \to 0, \qquad \text{as } \varepsilon \to 0. \tag{3}$$

Finally, we show that if at least one of the f_j's is undetectable, then

$$\liminf_{\varepsilon\to 0}\inf_{\tilde\eta}\sup_{\eta\in\mathcal{H}_{d,s}}\sup_{f\in\mathcal{F}^d_{s,\sigma}} s^{-1}\mathbf{E}_{f,\eta}|\tilde\eta - \eta| > 0, \tag{4}$$

that is, almost full recovery is impossible.

The selector η^* that satisfies (3) is said to provide asymptotically *almost full recovery* of a signal $f \in \mathcal{F}^d_{s,\sigma}$ in model (1); its maximum risk is small relative to the number of non-zero components. If, in addition, inequality (4) holds true, then the selection procedure based on η^* is the best possible (in the asymptotically minimax sense). The notion of optimality that we use is borrowed from the minimax hypothesis testing theory.

In the present setup, adaptive (in s) variable selection in high dimensions presents several challenges. First, one has to construct a good non-adaptive selector. Second, having that selector available, one has to adapt it to unknown values of the parameter s. It turns out that, when s is known, both exact and almost full recovery can be achieved by a suitably designed thresholding procedure (see Section 3.1 for details). The problem of adaptation of this procedure to unknown values of s was tackled and solved in [13], but in the case of exact recovery only. Handling the same problem in the almost full recovery case will bring us in this paper to the use of Lepski's method. This method was proposed for adaptive estimation in a Gaussian white noise model. The reason why adaptive reconstruction of most relevant components of f turns out to be more challenging than adaptive reconstruction of all components of f lies in the very nature of the thersholding procedure as defined in (20). In contrast to the exact selector given by (22) whose threshold is set regardless of the value of s, thresholding in (20) does depend on s.

The paper is organized as follows. In Section 2 we present some general results of the asymptotically minimax hypothesis testing theory and provide details on their use for the two function spaces of our interest. In Section 3 we translate the initial problem to the one in terms of the Fourier coefficients and, for both function spaces in hand, obtain almost full selectors for a known sparsity parameter s. In addition to that, we derive conditions under which almost full variable selection is possible. Adaptive selectors for the function spaces in hand are developed in Section 4. To complete the picture, we also introduce an adaptive selection procedure that gives exact reconstruction for the space of analytic functions. Our main results are stated in Section 4 and proved in Section 5.

2 The building blocks

As in [13], the recovery problem under study will be connected to that of hypothesis testing. Before stating and proving our main results, we shall discuss some important tools of minimax hypothesis testing that will be used in the subsequence sections. For a complete exposition of the subject, see [15] and the review papers [10, 11, 12].

2.1 Extreme problem for ellipsoids: general case

In asymptotically minimax hypothesis testing, when dealing with classes of smooth functions, the first common step is to transform the initial problem involving a class of functions to the corresponding problem in the space of Fourier coefficients. For this, let $\{\phi_k(x)\}_{k\in\mathbb{Z}}$ be the orthonormal basis in $L_2[0,1]$ given by

$$\phi_0(x) = 1, \quad \phi_k(x) = \sqrt{2}\cos(2\pi kx), \quad \phi_{-k}(x) = \sqrt{2}\sin(2\pi kx), \quad k > 0.$$

If $g \in L_2[0,1]$, then $g(x) = \sum_{k\in\mathbb{Z}} \theta_k\phi_k(x)$, where $\theta_k = (g, \phi_k)$ is the kth Fourier coefficient of g, and $\|g\|_2^2 = \sum_{k\in\mathbb{Z}} \theta_k^2$. Let $\mathcal{F}_\sigma$ be a function space depending on a parameter $\sigma > 0$ that is a subset of $L_2[0,1]$. Suppose that $g \in \mathcal{F}_\sigma \subset L_2[0,1]$ is observed in a univariate Gaussian white noise of intensity ε, and we wish to test the null hypothesis $H_0 : g = 0$ versus a sequence of alternatives $H_{1\varepsilon} : g \in \mathcal{F}_\sigma(r_\varepsilon)$, where the set $\mathcal{F}_\sigma(r_\varepsilon)$ is given by (2). For the two function spaces of interest, the norm of an element g is expressed as $\|g\|_\sigma^2 = \sum_{k\in\mathbb{Z}} c_k^2\theta_k^2$ with specified coefficients $c_k^2 = c_k^2(\sigma)$ (see formulas (8) and (12) below). In the sequence space of Fourier coefficients, the set $\mathcal{F}_\sigma(r_\varepsilon)$ corresponds to the ellipsoid in the space $l_2(\mathbb{Z})$ with semi-axis $c_k = c_k(\sigma)$ and a small neighbourhood of the point $\theta = 0$ removed:

$$\Theta_\sigma(r_\varepsilon) = \left\{\theta = (\theta_k)_{k\in\mathbb{Z}} \in l_2(\mathbb{Z}) : \sum_{k\in\mathbb{Z}} c_k^2\theta_k^2 \le 1, \sum_{k\in\mathbb{Z}} \theta_k^2 \ge r_\varepsilon^2\right\}. \tag{5}$$

For constructing an asymptotically almost full selector, we shall need some facts from the minimax theory of hypothesis testing. Denote by $\theta^*(r_\varepsilon) = (\theta_k^*(r_\varepsilon))_{k\in\mathbb{Z}}$ the solution to the extreme problem

$$\frac{1}{2\varepsilon^4}\sum_{k\in\mathbb{Z}} \theta_k^4 \to \inf_{\theta\in\Theta_\sigma(r_\varepsilon)}, \tag{6}$$

and let $u_\varepsilon^2(r_\varepsilon) = u_\varepsilon^2(\Theta_\sigma(r_\varepsilon))$ be the value of the problem, that is,

$$u_\varepsilon^2(r_\varepsilon) = \frac{1}{2\varepsilon^4}\inf_{\theta\in\Theta_\sigma(r_\varepsilon)}\sum_{k\in\mathbb{Z}} \theta_k^4 = \frac{1}{2\varepsilon^4}\sum_{k\in\mathbb{Z}} (\theta_k^*(r_\varepsilon))^4.$$

The function $u_\varepsilon^2(r_\varepsilon)$ plays a key role in the minimax theory of hypothesis testing. It controls the minimax total error probability and is used to set a cut-off point of the asymptotically minimax test procedure. The detection boundary r_ε^* in the problem of testing $H_0 : \theta = 0$ versus $H_1 : \theta \in \Theta_\sigma(r_\varepsilon)$ is determined by the relation $u_\varepsilon(r_\varepsilon^*) \asymp 1$. The function $u_\varepsilon(r_\varepsilon)$ is a non-decreasing function of the argument r_ε which possesses a kind of 'continuity' property. Namely, for any $\epsilon > 0$ there exist $\Delta > 0$ and $\varepsilon_0 > 0$ such that for any $\delta \in (0, \Delta)$ and $\varepsilon \in (0, \varepsilon_0)$,

$$u_\varepsilon(r_\varepsilon) \le u_\varepsilon((1+\delta)r_\varepsilon) \le (1+\epsilon)u_\varepsilon(r_\varepsilon). \tag{7}$$

These and some other facts about $u_\varepsilon^2(r_\varepsilon)$ can be found in [10, Sec. 3.2] and [15, Sec. 5.2.3]).

For standard function spaces with the norm $\|g\|_\sigma$ defined (under the periodic constraints) in terms of Fourier coefficients as $\|g\|_\sigma^2 = \sum_{k\in\mathbb{Z}} \theta_k^2 c_k^2$, the form of the extremal sequence $(\theta_k^*(r_\varepsilon)_{k\in\mathbb{Z}}$ in problem (6) as well as the sharp asymptotics for $u_\varepsilon(r_\varepsilon)$ are available. Below we cite some relevant results for the two function spaces $\mathcal{F}_\sigma$ of our interest: the Sobolev space of periodic σ-smooth function on $\mathbb{R}$ and the space of periodic functions on $\mathbb{R}$ that admit an analytic continuation to the strip around the real line.

2.2 Extreme problem for Sobolev ellipsoids

Let $\mathcal{F}_\sigma$ with $\sigma > 0$ denote the Sobolev space of σ-smooth 1-periodic functions on $\mathbb{R}$. Define the norm $\|\cdot\|_\sigma$ on $\mathcal{F}_\sigma$ by the formula

$$\|f\|_\sigma^2 = \sum_{k\in\mathbb{Z}} \theta_k^2 c_k^2, \quad c_k^2 = c_k^2(\sigma) = (2\pi|k|)^{2\sigma}, \tag{8}$$

where θ_k is the kth Fourier coefficient of f with respect to $\{\phi_k(x)\}_{k\in\mathbb{Z}}$. If σ is an integer, then under the periodic constraints (when the function admits 1-periodic $[\sigma]$-smooth extension on the real line) the norm as in (8) corresponds to

$$\|f\|_\sigma^2 = \int_0^1 \left(f^{(\sigma)}(x)\right)^2 dx.$$

For a function $f \in \mathcal{F}_\sigma$ consider testing the hypothesis $H_0 : f = 0$ versus the alternative $H_1 : f \in \mathcal{F}_\sigma(r_\varepsilon)$, where for a positive family $r_\varepsilon \to 0$

$$\mathcal{F}_\sigma(r_\varepsilon) = \{f \in \mathcal{F}_\sigma : \|f\|_\sigma \le 1, \ \|f\|_2 \ge r_\varepsilon\}.$$

Switching from Sobolev balls $\{f \in \mathcal{F}_\sigma : \|f\|_\sigma \le 1\}$ to Sobolev ellipsoids $\{\theta \in l_2(\mathbb{Z}) : \sum_{k\in\mathbb{Z}} c_k^2\theta_k^2 \le 1\}$ leads to the problem of testing $H_0 : \theta = 0$ versus $H_1 : \theta \in \Theta_\sigma(r_\varepsilon)$. The test procedure that does the best in distinguishing between the latter two hypotheses is obtained by solving the extreme problem (6) with the semi-axes c_k defined as in (8); see Section 3 of [10] for details. The extremal sequence $(\theta_k^*(r_\varepsilon)_{k\in\mathbb{Z}}$ satisfies (see, for example, [10, § 3.2] and Theorem 2 in [16]):

$$(\theta_k^*(r_\varepsilon))^2 \asymp r_\varepsilon^{2+1/\sigma}\left(1 - (2\pi|k|/K_\varepsilon)^{2\sigma}\right) \ \text{ for } \ 1 \le |k| \le K_\varepsilon \quad \text{and} \quad \theta_k^*(r_\varepsilon) = 0 \ \text{ otherwise}, \tag{9}$$

where

$$K_\varepsilon = \lfloor (4\sigma+1)^{1/(2\sigma)} r_\varepsilon^{-1/\sigma} \rfloor. \tag{10}$$

The sharp asymptotics for $u_\varepsilon(r_\varepsilon)$ are of the form (see [15, § 4.3.2] and Theorems 2 and 4 in [16])

$$u_\varepsilon(r_\varepsilon) \sim C(\sigma) r_\varepsilon^{2+1/(2\sigma)} \varepsilon^{-2}, \quad \varepsilon \to 0, \tag{11}$$

where (see, for example, p. 104 of [10])

$$C(\sigma) = 2\sigma \left[\left(1 + \frac{1}{4\sigma}\right)(1+4\sigma)^{1/(2\sigma)} \left(B\left(\frac{1}{2\sigma}, 2\right)\right)^{1/\sigma}\right]^{-1},$$

and $B(\cdot,\cdot)$ is the Euler beta-function.

2.3 Extreme problem for the ellipsoids of analytic functions

The following example of $\mathcal{F}_\sigma$ is also well known in nonparametric estimation and hypothesis testing. Let $\mathcal{F}_\sigma$ with $\sigma > 0$ be the class of 1-periodic functions f on $\mathbb{R}$ admitting a continuation to the strip $S_\sigma = \{z = x + iy : |y| \leq \sigma\} \subset \mathbb{C}$ such that $f(x + iy)$ is analytic on the interior of S_σ, bounded on S_σ and

$$\int_0^1 |f(x \pm i\sigma)|^2 \, dx < \infty.$$

Let the norm $\|\cdot\|_{1,\sigma}$ on $\mathcal{F}_\sigma$ be given by (see, for example, [7])

$$\|f\|_{1,\sigma}^2 = \int_0^1 \left(\mathrm{Re} f(x + i\sigma)\right)^2 \, dx.$$

In terms of the Fourier coefficients, the squared norm $\|f\|_{1,\sigma}^2$ takes the form

$$\|f\|_{1,\sigma}^2 = \sum_{k \in \mathbb{Z}} \theta_k^2 c_k^2, \quad c_k^2 = c_k^2(\sigma) = \cosh^2(2\pi\sigma k).$$

In view of the relations

$$\exp(|x|) \leq 2\cosh(x) \leq 2\exp(|x|), \quad x \in \mathbb{R},$$

we may also consider an equivalent norm $\|\cdot\|_\sigma$ defined as

$$\|f\|_\sigma^2 = \sum_{k \in \mathbb{Z}} \theta_k^2 c_k^2, \quad c_k = c_k(\sigma) = \exp(2\pi\sigma|k|). \tag{12}$$

We have chosen to deal with the latter norm as it is easier to study.

The ball $\{f \in \mathcal{F}_\sigma : \|f\|_\sigma \leq 1\}$ corresponds to the ellipsoid $\{\theta \in l_2(\mathbb{Z}) : \sum_{k \in \mathbb{Z}} c_k^2 \theta_k^2 \leq 1\}$ with the semi-axes c_k defined as in (12). Thus translating the problem of testing $H_0 : f = 0$ versus $H_1 : f \in \mathcal{F}_\sigma(r_\varepsilon)$ to the one in terms of Fourier coefficients brings us to testing $H_0 : \theta = 0$ versus $H_1 : \theta \in \Theta_\sigma(r_\varepsilon)$. The asymptotically minimax test procedure that distinguishes between these two hypotheses is obtained by solving the extreme problem (6) with the semi-axes c_k defined as in (12). The elements of the extremal sequence $(\theta_k^*(r_\varepsilon)_{k \in \mathbb{Z}}$ in problem (6) with the semi-axis c_k as above may be taken as constants (independent of k) satisfying as $\varepsilon \to 0$ (see, for example, Section 3 in [10])

$$\theta_k^*(r_\varepsilon) \asymp r_\varepsilon \log^{-1/2}(r_\varepsilon^{-1}) \ \text{ for } \ 1 \leq |k| \leq K_\varepsilon \quad \text{and} \quad \theta_k^*(r_\varepsilon) = 0 \ \text{ otherwise}, \tag{13}$$

where

$$K_\varepsilon = \lfloor (2\pi\sigma)^{-1} \log(r_\varepsilon^{-1}) \rfloor, \tag{14}$$

and we have

$$u_\varepsilon(r_\varepsilon) \sim \left(\frac{r_\varepsilon}{\varepsilon}\right)^2 \frac{(2\pi\sigma)^{1/2}}{\log^{1/2}(r_\varepsilon^{-1})}. \tag{15}$$

Formulas (13)–(15), as well as formulas (9)–(11), will be employed to construct almost full selectors for the two function spaces under study.

3 Variable selection in a sequence space model

By sufficiency, the problem of recovering f observed in the Gaussian white noise model can be transformed to an equivalent problem in a sequence space model. Acting as in [13], for the index $l \in \mathbb{Z}^d$ whose jth component is equal to k and the other components are equal to zero, define the function

$$\phi_{j,k}(\mathbf{x}) = \phi_l(\mathbf{x}) = \phi_k(x_j), \quad \mathbf{x} = (x_1, \dots, x_d) \in [0,1]^d, \quad 1 \le j \le d, \quad k \in \mathbb{Z},$$

and denote by $\theta_{j,k} = (f, \phi_{j,k}) = \int_0^1 \phi_k(x) f_j(x)\, dx$ the kth Fourier coefficient of the jth component f_j of a signal $f = \sum_{j=1}^d \eta_j f_j$. Consider the sequence space model

$$X_{j,k} = \eta_j \theta_{j,k} + \varepsilon \xi_{j,k}, \quad \xi_{j,k} \overset{i.i.d.}{\sim} N(0,1), \quad 1 \le j \le d, \quad k \in \mathbb{Z}, \tag{16}$$

where $X_{j,k} = X_\varepsilon(\phi_{j,k})$ are the empirical Fourier coefficients and the collection $(\eta_1\theta_1, \dots, \eta_d\theta_d)$ consists of sequences $\eta_j\theta_j = (\eta_j\theta_{j,k})_{k\in\mathbb{Z}}$ such that $(\eta_j) \in \mathcal{H}_{d,s}$ and for all $1 \le j \le d$,

$$\theta_j = (\theta_{j,k}) \in l_2(\mathbb{Z}), \quad \sum_{k\in\mathbb{Z}} c_k^2 \theta_{j,k}^2 \le 1. \tag{17}$$

In this paper we have chosen to deal with the latter model, which is technically more convenient. Although the set of θ_js in (17) involves an orthogonal system in L_2^d, the results on minimax errors and risks do not depend on the choice of this orthogonal system because the random variables $X_{j,k}$, which generate a sufficient σ-algebra for $f \in \mathcal{F}_{s,\sigma}^d$, are independent normal $N(\eta_j\theta_{j,k}, \varepsilon^2)$. Thus the distribution of $\{X_{j,k}\}$ depends on the Fourier coefficients $\theta_{j,k}$ of f with respect to the system $\{\phi_{j,k}\}$ but not on the choice of $\{\phi_{j,k}\}$. Using a suitable finite collection of the random variables $X_{j,k}$ as defined in (16), we wish to construct an optimal selection procedure that recovers most non-zero components of $(\eta_1\theta_1, \dots, \eta_d\theta_d)$, but not all of them.

3.1 Almost full variable selection in the non-adaptive case

We first consider a non-adaptive setup when the sparsity parameter s is known. When dealing with the problem of variable selection in model (16), we make use of the statistics, cf. asymptotically minimax test statistics in Section 3.1 of [10],

$$t_j = t_j(s) = \sum_{1\le|k|\le K_\varepsilon} \omega_k(r_\varepsilon^*(s)) \left[\left(\frac{X_{j,k}}{\varepsilon} \right)^2 - 1 \right], \quad j = 1, \dots, d, \tag{18}$$

where for any $r_\varepsilon > 0$ the weight functions $\omega_k(r_\varepsilon)$ are given by the formula

$$\omega_k(r_\varepsilon) = \frac{1}{2\varepsilon^2} \frac{(\theta_k^*(r_\varepsilon))^2}{u_\varepsilon(r_\varepsilon)}, \quad 1 \le |k| \le K_\varepsilon,$$

and the number $r_\varepsilon^*(s) > 0$ is the solution of the equations

$$\frac{u_\varepsilon(r_\varepsilon^*(s))}{\sqrt{2\log(d/s)}} = 1. \tag{19}$$

For both function spaces of interest, the quantities K_ε, $\theta_k^*(r_\varepsilon)$, and $u_\varepsilon(r_\varepsilon)$ in formula (18) are specified in Section 2. The sparsity parameter $s \in \{1, 2, \dots, d\}$ is assumed to be small relative to d, that is, $s = o(d)$. Note that the weights $\omega_k(r_\varepsilon)$ are normalized to have $\sum\limits_{1\le|k|\le K_\varepsilon} \omega_k^2(r_\varepsilon) = 1/2$.

Now we define a non-adaptive *almost full selector* to be

$$\check{\eta} = (\check{\eta}_1, \dots, \check{\eta}_d), \quad \check{\eta}_j = \mathbb{I}\left(t_j > \sqrt{2\log(d/s) + \delta \log d}\right), \quad j = 1, \dots, d, \tag{20}$$

where $\delta = \delta_\varepsilon > 0$ satisfies

$$\delta \to 0 \quad \text{and} \quad \delta \log d \to \infty, \quad \text{as } \varepsilon \to 0. \tag{21}$$

The arguments as in the proof of Theorem 1 show that for Sobolev ellipsoids, under the conditions, cf. (23),

$$\log d = o(\varepsilon^{-2/(2\sigma+1)}), \quad \liminf_{\varepsilon \to 0} \frac{u_\varepsilon(r_\varepsilon)}{\sqrt{2\log(d/s)}} > 1,$$

the selector $\check{\eta}$ reconstructs almost all relevant components of a vector $\eta \in \mathcal{H}_{d,s}$, and hence asymptotically provides almost full recovery of a signal $f \in \mathcal{F}^d_{s,\sigma}$ in model (1).

To illustrate the difference between exact and almost full reconstruction in adaptive settings, assume that $\mathcal{F}_\sigma$ is the Sobolev space. In this case, a selector (see Section 3.1 of [13] with s in place of $d^{1-\beta}$)

$$\eta^* = (\eta^*_1, \dots, \eta^*_d), \quad \eta^*_j = \mathbb{I}\left(t^*_j > \sqrt{(2+\delta)\log d}\right), \quad j = 1, \dots, d, \tag{22}$$

where the statistics t^*_j are defined similar to the t_j as in (18) with the relation

$$\frac{u_\varepsilon(r^*_\varepsilon(s))}{\sqrt{2\log d} + \sqrt{2\log s}} = 1$$

instead of (19), turns our to be a non-adaptive *exact selector*, as long as

$$\log d = o(\varepsilon^{-2/(2\sigma+1)}) \quad \text{and} \quad \liminf_{\varepsilon \to 0} \frac{u_\varepsilon(r_\varepsilon)}{\sqrt{2\log d} + \sqrt{2\log s}} > 1. \tag{23}$$

Under the above conditions, the procedure based on η^* selects correctly all non-zero components of a vector $\eta \in \mathcal{H}_{d,s}$, and hence provides exact recovery of a signal $f \in \mathcal{F}^d_{s,\sigma}$ in model (1).

Contrasting with formula (22), the threshold in (20) is set at a lower level and is dependent on the parameter s. The latter fact makes the idea of adaption suggested in [13] for the exact reconstruction case invalid (see Section 3.3. for details). In the next section, we obtain the desired adaptive selector by using Lepski's method. Before doing that, we provide conditions on d as a function of ε under which the thresholding procedure (20), as well as its adaptive version introduced in Section 4.1, gives asymptotically almost full reconstruction of a function $f \in \mathcal{F}^d_{s,\sigma}$.

3.2 Conditions for almost full variable selection

Consider now the question of determining conditions on d as a function of ε under which almost full variable selection is possible. Violation of these conditions will lead to entirely different selection strategies.

In the sequence space of Fourier coefficients, consider testing the null hypothesis $H_{0j} : \theta_j = 0$ versus the alternative $H_{1j} : \theta_j \in \Theta_\sigma(r_\varepsilon)$, where the set $\Theta_\sigma(r_\varepsilon)$ is given by (5). It is easy to see that under the null hypothesis H_{0j}, we have (see, for example, Section 4.1 of [13])

$$\mathbf{E}_0(t_j) = 0, \quad \mathbf{Var}_0(t_j) = 1,$$

while under the alternative H_{1j}, where for all sufficiently small ε a small parameter $r_\varepsilon > 0$ satisfies $r_\varepsilon / r^*_\varepsilon(s) > 1$,

$$\begin{aligned} \mathbf{E}_{\theta_j}(t_j) &= \varepsilon^{-2} \sum_{1 \le |k| \le K_\varepsilon} \omega_k(r^*_\varepsilon(s)) \theta^2_{j,k} \ge u_\varepsilon(r^*_\varepsilon(s)), \qquad (24) \\ \mathbf{Var}_{\theta_j}(t_j) &= 1 + O(\mathbf{E}_{\theta_j}(t_j) \max_{1 \le |k| \le K_\varepsilon} \omega_k(r^*_\varepsilon(s))). \end{aligned}$$

Furthermore, under the above restrictions on r_ε and $r^*_\varepsilon(s)$ the following result holds (in case of Sobolev spaces, see Proposition 7.1 in [5] and Lemma 1 in [13]; in case of the space $\mathcal{F}_\sigma$ of analytic functions, the proof is similar to that of Sobolev spaces).

Let the quantity $T = T_\varepsilon \to -\infty$ and the weight functions $\omega_k(r^*_\varepsilon(s))$ as in (18) be such that as $\varepsilon \to 0$

$$T \max_{1 \le |k| \le K_\varepsilon} \omega_k(r^*_\varepsilon(s)) \to 0 \quad \text{and} \quad \mathbf{E}_{\theta_j}(t_j(s)) \max_{1 \le |k| \le K_\varepsilon} \omega_k(r^*_\varepsilon(s)) \to 0. \qquad (25)$$

Then as $\varepsilon \to 0$

$$\mathbf{P}_0(t_j \le T) \le \exp\left(-\frac{T^2}{2}(1 + o(1))\right), \qquad (26)$$

and for all $j = 1, \dots, d$, uniformly in $\theta_j \in \Theta_\sigma(r_\varepsilon)$,

$$\mathbf{P}_{\theta_j}(t_j - \mathbf{E}_{\theta_j}(t_j) \le T) \le \exp\left(-\frac{T^2}{2}(1 + o(1))\right). \qquad (27)$$

For both function spaces $\mathcal{F}_\sigma$ of our interest, the exponential bounds (26) and (27) will be applied below to the quantity $T = T_\varepsilon \to -\infty$ of order $O(\log^{1/2} d)$. This observation and assumption (19) transform requirement (25) into

$$\log^{1/2} d \max_{1 \le |k| \le K_\varepsilon} \omega_k(r^*_\varepsilon(s)) \to 0, \quad \varepsilon \to 0, \qquad (28)$$

Condition (28) gives a restriction on the growth of $d = d_\varepsilon$ ensuring that the selection procedure works as designed. Indeed, as shown in Section 4.1 in [13], for the Sobolev space of σ-smooth functions, one has

$$\omega_k(r_\varepsilon) \asymp r_\varepsilon^{1/(2\sigma)} \ \text{ for } \ 1 \le |k| \le K_\varepsilon,$$

and

$$r^*_\varepsilon(s) \asymp (\varepsilon \log d)^{\sigma/(4\sigma+1)}.$$

Therefore condition (28) is fulfilled when

$$\log d = o(\varepsilon^{-2/(2\sigma+1)}) \qquad (29)$$

In case of the space $\mathcal{F}_\sigma$ of analytic functions, one has

$$\omega_k(r_\varepsilon) \asymp \log^{-1/2}(r_\varepsilon^{-1}) \ \text{ for } \ 1 \le |k| \le K_\varepsilon, \tag{30}$$

and, in view of (11) and (19), the quantity $r_\varepsilon^*(s)$ satisfies

$$\log^{1/2} d \asymp \left(\frac{r_\varepsilon^*(s)}{\varepsilon}\right)^2 \log^{-1/2}\left((r_\varepsilon^*(s))^{-1}\right),$$

implying

$$r_\varepsilon^*(s) \asymp \varepsilon \log^{1/4}(d) \log^{1/4}((r_\varepsilon^*(s))^{-1}).$$

Therefore $\log\left((r_\varepsilon^*(s))^{-1}\right) \sim \log(\varepsilon^{-1})$, and (see (30))

$$\omega_k(r_\varepsilon^*(s)) \asymp \log^{-1/2}(\varepsilon^{-1}).$$

From this, the technical condition (28) holds true when, cf. formula (30),

$$\log d = o(\log(\varepsilon^{-1})), \quad \varepsilon \to 0. \tag{31}$$

4 Main results

In this section, we consider a more realistic problem when the sparsity parameter s is *unknown*. We derive conditions under which almost full variable selection is possible, and construct a selector for which the Hamming distance is much smaller than the number of relevant components (see Theorems 3 and 5). Our selector is adaptive in the sparsity parameter s and is unimprovable in the asymptotically minimax sense (see Theorems 4 and 6). In addition to that, in Section 4.2 we provide asymptotically exact selection procedure for the space of analytic functions that is adaptive in the sparsity parameter s.

4.1 Almost full variable selection in the adaptive case

In this subsection, the selector $\check{\eta}$ as in (20) will be used to obtain the corresponding adaptive procedure. To avoid losses due to adaptation, we will have to limit the range of the possible values of s. Namely, we assume that for some constants $0 < c < C < 1$

$$c \le \liminf_{d\to\infty}(\log s/\log d) \le \limsup_{d\to\infty}(\log s/\log d) \le C, \tag{32}$$

and define the set

$$\mathcal{S}_d = \{s \in \{1, \dots, d\} \text{ is such that condition (32) holds}\}$$

over which the adaptive selector that we propose yields almost full selection. The restriction on s as in (32) is relatively mild. For instance, any $s = d^{1-\beta}$ with $\beta \in [b, B]$ for some constants $0 < b < B < 1$ belongs to $\mathcal{S}_d$.

To construct the desired selector, for some $\Delta = \Delta_d > 0$ and $M = \lceil (C-c)/\Delta \rceil + 1$, pick grid points over the interval $(1, d)$:

$$s_1 = d^c, \quad s_m = s_{m-1}d^\Delta = s_1 d^{(m-1)\Delta}, \quad 2 \le m \le M, \tag{33}$$

and assume that

$$\Delta \to 0, \quad \Delta \log d \to 0, \quad \text{as } \; d \to \infty, \tag{34}$$

yielding $d^{\Delta} \le \text{const}$ for all large enough d. For each $m = 1, \ldots, M$, let the parameter $r_\varepsilon^*(s_m) > 0$ be determined by the equation, cf. (19),

$$\frac{u_\varepsilon(r_\varepsilon^*(s_m))}{\sqrt{2\log(d/s_m)}} = 1,$$

where, depending on a type of the ellipsoid $\Theta_\sigma(r_\varepsilon)$ we are dealing with, the function $u_\varepsilon(r_\varepsilon)$ satisfies either (11) or (15).

Similar to the case of known s, consider weighted chi-square type statistics, cf. (18),

$$t_j(s_m) = \sum_{1\le|k|\le K_\varepsilon} \omega_k(r_\varepsilon^*(s_m))\Big[\Big(\frac{X_{j,k}}{\varepsilon}\Big)^2 - 1\Big], \quad j = 1, \ldots, d, \quad m = 1, \ldots, M.$$

with weight functions

$$\omega_k(r_\varepsilon^*(s_m)) = \frac{1}{2\varepsilon^2}\frac{(\theta_k^*(r_\varepsilon^*(s_m)))^2}{u_\varepsilon(r_\varepsilon^*(s_m))}, \quad 1 \le |k| \le K_\varepsilon,$$

possessing the property $\sum_{1\le|k|\le K_\varepsilon} \omega_k^2(r_\varepsilon^*(s_m)) = 1/2$. The values of $\theta_k^*(r_\varepsilon^*(s))$ and K_ε depend on the function space under consideration. For the Sobolev space in hand, $\theta_k^*(r_\varepsilon^*(s))$ and K_ε are as in (9) and (10); for the space of analytic functions, $\theta_k^*(r_\varepsilon^*(s))$ and K_ε are as in (13) and (14).

Next, for all $j = 1, \ldots, d$ and $m = 1, \ldots, M$, set

$$\widehat{\eta}_j(s_m) = \mathbb{I}\left(t_j(s_m) > \sqrt{2\log(d/s_m) + \delta \log d}\right),$$

where $\delta = \delta_\varepsilon > 0$ satisfies (21), and define an adaptive selector of a vector $\eta \in \mathcal{H}_{d,s}$ by the formula

$$\widehat{\eta}(s_{\widehat{m}}) = (\widehat{\eta}_1(s_{\widehat{m}}), \ldots, \widehat{\eta}_d(s_{\widehat{m}})), \tag{35}$$

where $\widehat{m}$ is chosen by Lepski's method (see Section 2 of [17]) as follows:

$$\widehat{m} = \min\left\{1 \le m \le M \,:\, |\widehat{\eta}(s_m) - \widehat{\eta}(s_i)| \le v_i \;\text{ for all }\; i \ge m\right\}.$$

Here the quantities $v_i = v_{i,d}$ are set to be

$$v_i = s_i/\tau_d, \quad m \le i \le M,$$

with a sequence of numbers $\tau_d \to \infty$ satisfying (recall that $d = d_\varepsilon \to \infty$ as $\varepsilon \to 0$)

$$\tau_d = o\left(\min(\log d, d^{\delta/2})\right), \quad \text{as } \; \varepsilon \to 0.$$

Algorithmically, Lepski's procedure for choosing $\hat{m}$ works as follows. We start by setting $\hat{m} = M$ and attempt to decrease the value of $\hat{m}$ from M to $M-1$. If $|\widehat{\eta}(s_{M-1}) - \widehat{\eta}(s_M)| \le v_M$, we set $\hat{m} = M-1$; otherwise, we keep $\hat{m}$ equal to M. In case $\hat{m}$ is decreased to $M-1$, we continue the process attempting to decrease it further. If $|\widehat{\eta}(s_{M-2}) - \widehat{\eta}(s_{M-1})| \le v_{M-1}$ and $|\widehat{\eta}(s_{M-2}) - \widehat{\eta}(s_M)| \le v_M$, we set $\hat{m} = M-2$; otherwise, we keep $\hat{m}$ equal to $M-1$; and so on. Notice that by construction $v_M \ge v_{M-1} \ge \ldots \ge v_1$.

4.2 Exact variable selection for analytic functions in the adaptive case

The problem of adaptive reconstruction of sparse additive functions in the Gaussian white noise model was studied in the only case of σ-smooth functions, see [13]. Before handling the problem of almost full variable selection in adaptive settings, we complement the findings in [13] by presenting an adaptive exact selector for the space of analytic functions. The strategy is similar to the one suggested in [13] for σ-smooth functions, but the parameters of the statistics and the condition on the dimension d are different.

Consider a sequence space model that corresponds to the Gaussian white noise model with f from the class of analytic functions $\mathcal{F}_\sigma$ as defined in Section 2.3. Let $1 < s_1 < s_1 < \ldots < s_M < d$ be the grid of points as in (33). For any $m = 1, \ldots, M$, let the parameter $r^*_{\varepsilon,m} > 0$ be determined by the equation

$$\frac{u_\varepsilon(r^*_{\varepsilon,m})}{\sqrt{2\log d} + \sqrt{2\log s_m}} = 1.$$

Consider weighted chi-square type statistics

$$t_{j,m} = \sum_{1\le|k|\le K_\varepsilon} \omega_k(r^*_{\varepsilon,m})\left[\left(\frac{X_{j,k}}{\varepsilon}\right)^2 - 1\right], \quad j = 1, \ldots, d, \quad m = 1, \ldots, M.$$

with weight functions

$$\omega_k(r^*_{\varepsilon,m}) = \frac{1}{2\varepsilon^2}\frac{(\theta^*_k(r^*_{\varepsilon,m}))^2}{u_\varepsilon(r^*_{\varepsilon,m})}$$

obeying the normalization condition $\sum_{k\in\mathbb{Z}} \omega^2_k(r^*_{\varepsilon,m}) = 1/2$. Next, for all $j = 1, \ldots, d$ and $m = 1, \ldots, M$, set

$$\eta_{j,m} = \mathbb{I}\left(t_{j,m} > \sqrt{(2+\delta)(\log d + \log M)}\right),$$

and define an *adaptive exact selector* η^{**} of a vector $\eta \in \mathcal{H}_{d,s}$ by the formula (see formula (18) in [13])

$$\eta^{**} = (\eta^{**}_1, \ldots, \eta^{**}_d), \quad \eta^{**}_j = \max_{1\le m\le M} \eta_{j,m}, \quad j = 1, \ldots, d. \tag{36}$$

The idea behind the selector η^{**} is as follows. The jth component of a signal is viewed active if at least one of the statistics $t_{j,m}$, $m = 1, \ldots, M$, detects it. Therefore, thinking of $\eta_{j,m}$ and η^{**}_j as test functions, we get that the probability of having θ_j incorrectly undetected does not exceed the respective probability with the $\eta_{j,m}$ test, where s_m is close to the true (but unknown) value of s. Furthermore, the probability that η^{**}_j incorrectly detects θ_j is less than the sum of the respective probabilities for the $\eta_{j,m}$ tests over all $m = 1, \ldots, M$, and is small by the choice of threshold.

Let the set $\Theta_{\sigma,d}(r_\varepsilon)$ be as in (37) with the coefficients c_k given by (12). The following two theorems, whose proofs are similar to those of Theorems 3 and 4 in [13], hold true.

Theorem 1. *Let $s \in \{1, \ldots, d\}$ be such that $s = o(d)$. Assume that $\log d = o(\log \varepsilon^{-1})$ and that the quantity $r_\varepsilon = r_\varepsilon(s) > 0$ satisfies*

$$\liminf_{\varepsilon\to 0} \frac{u_\varepsilon(r_\varepsilon)}{\sqrt{2\log d} + \sqrt{2\log s}} > 1.$$

Then as $\varepsilon \to 0$

$$\sup_{\eta \in \mathcal{H}_{d,s}} \sup_{\theta \in \Theta_{\sigma,d}(r_\varepsilon)} \mathbf{E}_{\eta,\theta} |\eta - \eta^{**}| \to 0,$$

where η^{**} is the selector of vector η as defined in (36)*.*

Theorem 2. *Let $s \in \{1, \ldots, d\}$ be such that $s = o(d)$. Assume that $\log d = o(\log \varepsilon^{-1})$ and that the quantity $r_\varepsilon = r_\varepsilon(s) > 0$ satisfies*

$$\limsup_{\varepsilon \to 0} \frac{u_\varepsilon(r_\varepsilon)}{\sqrt{2\log d + \sqrt{2 \log s}}} < 1.$$

Then

$$\liminf_{\varepsilon \to 0} \inf_{\tilde{\eta}} \sup_{\eta \in \mathcal{H}_{d,s}} \sup_{\theta \in \Theta_{\sigma,d}(r_\varepsilon)} \mathbf{E}_{\eta,\theta} |\eta - \tilde{\eta}| > 0,$$

where the infimum is over all selectors $\tilde{\eta}$ of vector η in model (16)*.*

Remark 1. The sharp detection boundary in Theorems 1 and 2 which makes it possible to decide on whether we are in a position to proceed further with variable selection or not, is determined in terms of the function $u_\varepsilon(r_\varepsilon)$ with sharp asymptotics as in (11) and (15). The use of $u_\varepsilon(r_\varepsilon)$ instead of r_ε makes it easy to build a bridge between variable selection in Gaussian white noise setting and variable selection in regression setting as studied in Sec. 4 of [6]. In addition, using $u_\varepsilon(r_\varepsilon)$ instead of r_ε makes the statement of detectability condition precise. By 'continuity' of $u_\varepsilon(r_\varepsilon)$ as cited in (7), the conditions of Theorems 1 and 2 that separate detectible components from undetectable ones can be written in a usual form $\liminf_{\varepsilon \to 0} r_\varepsilon / r_\varepsilon^* > 1$ and $\limsup_{\varepsilon \to 0} r_\varepsilon / r_\varepsilon^* < 1$, where for Sobolev ellipsoids the sharp detection boundary r_ε^* is found explicitly from (11), and for the ellipsoids of analytic functions it is found implicitly from (15). Similar remark applies to Theorems 3 to 6 stated in Section 4.3 and 4.4,

4.3 Almost full variable selection for Sobolev balls

Consider the set $\Theta_\sigma(r_\varepsilon)$ as in (5) with the coefficients c_k given by (8), and define the set

$$\Theta_{\sigma,d}(r_\varepsilon) = \left\{ \theta = (\theta_j) : \theta_j = (\theta_{j,k}) \in l_2(\mathbb{Z}),\ \sum_{k \in \mathbb{Z}} c_k^2 \theta_{j,k}^2 \le 1,\ \sum_{k \in \mathbb{Z}} \theta_{j,k}^2 \ge r_\varepsilon^2,\ 1 \le j \le d \right\}. \tag{37}$$

Let $\widehat{\eta}(s_{\widehat{m}})$ be the selector given by (35) based on the statistics $t_j(s_m)$ as in (18), where the quantities $\theta_k^*(r_\varepsilon)$, K_ε, and $u_\varepsilon(r_\varepsilon)$ are specified by formulas (9), (10), and (11), respectively. The following theorem holds.

Theorem 3. *Let $s \in \{1, \ldots, d\}$ be such that* (32) *holds true. Assume that $\log d = o(\varepsilon^{-2/(2\sigma+1)})$ and that the quantity $r_\varepsilon = r_\varepsilon(s) > 0$ satisfies*

$$\liminf_{\varepsilon \to 0} \frac{u_\varepsilon(r_\varepsilon)}{\sqrt{2 \log(d/s)}} > 1.$$

Then as $\varepsilon \to 0$

$$\sup_{s \in \mathcal{S}_d} \sup_{\eta \in \mathcal{H}_{d,s}} \sup_{\theta \in \Theta_{\sigma,d}(r_\varepsilon)} s^{-1} \mathbf{E}_{\eta,\theta} |\widehat{\eta}(s_{\widehat{m}}) - \eta| \to 0.$$

Theorem 3 says that if all the hypotheses $H_{0j} : \theta_j \equiv 0$ and $H_{1j} : \theta_j \in \Theta_\sigma(r_\varepsilon)$, $j = 1, \dots, d$, separate asymptotically, then the selection procedure based on $\widehat{\eta}(s_{\widehat{m}})$ reconstructs almost all non-zero components of a vector $\eta \in \mathcal{H}_{d,s}$, and thus provides almost full recovery of $(\eta_1\theta_1, \dots, \eta_d\theta_d)$, uniformly in $\mathcal{S}_d$, $\mathcal{H}_{d,s}$, and $\Theta_{\sigma,d}(r_\varepsilon)$.

The next result shows that if the detectability condition is not met, almost full selection is impossible.

Theorem 4. *Let $s \in \{1, \dots, d\}$ be such that $s = o(d)$. Assume that $\log d = o(\varepsilon^{-2/(2\sigma+1)})$ and that the quantity $r_\varepsilon = r_\varepsilon(s) > 0$ satisfies*

$$\limsup_{\varepsilon \to 0} \frac{u_\varepsilon(r_\varepsilon)}{\sqrt{2\log(d/s)}} < 1.$$

Then

$$\liminf_{\varepsilon \to 0} \inf_{\tilde{\eta}} \sup_{\eta \in \mathcal{H}_{d,s}} \sup_{\theta \in \Theta_{\sigma,d}(r_\varepsilon)} s^{-1}\mathbf{E}_{\eta,\theta}|\eta - \tilde{\eta}| > 0,$$

where the infimum is over all selectors $\tilde{\eta}$ of a vector η in model (16).

4.4 Almost full variable selection for analytic functions

The results similar to Theorems 3 and 4 hold true for the space of analytic functions. Namely, consider the sets $\Theta_\sigma(r_\varepsilon)$ and $\Theta_{\sigma,d}(r_\varepsilon)$ as in (5) and (37) with the coefficients c_k given by (12). Again, let $\widehat{\eta}(s_{\widehat{m}})$ be the selector defined by (35) based on the statistics $t_j(s_m)$ as in (18), but the quantities $\theta_k^*(r_\varepsilon)$, K_ε, and $u_\varepsilon(r_\varepsilon)$ are now as in (13), (14), and (15), respectively. The following results hold true.

Theorem 5. *Let $s \in \{1, \dots, d\}$ be such that* (32) *holds true. Assume that $\log d = o(\log \varepsilon^{-1})$ and that the quantity $r_\varepsilon = r_\varepsilon(s) > 0$ satisfies*

$$\liminf_{\varepsilon \to 0} \frac{u_\varepsilon(r_\varepsilon)}{\sqrt{2\log(d/s)}} > 1.$$

Then as $\varepsilon \to 0$

$$\sup_{s \in \mathcal{S}_d} \sup_{\eta \in \mathcal{H}_{d,s}} \sup_{\theta \in \Theta_{\sigma,d}(r_\varepsilon)} s^{-1}\mathbf{E}_{\eta,\theta}|\widehat{\eta}(s_{\widehat{m}}) - \eta| \to 0.$$

Theorem 6. *Let $s \in \{1, \dots, d\}$ be such that $s = o(d)$. Assume that $\log d = o(\log \varepsilon^{-1})$ and that the quantity $r_\varepsilon = r_\varepsilon(s) > 0$ satisfies*

$$\limsup_{\varepsilon \to 0} \frac{u_\varepsilon(r_\varepsilon)}{\sqrt{2\log(d/s)}} < 1.$$

Then

$$\liminf_{\varepsilon \to 0} \inf_{\tilde{\eta}} \sup_{\eta \in \mathcal{H}_{d,s}} \sup_{\theta \in \Theta_{\sigma,d}(r_\varepsilon)} s^{-1}\mathbf{E}_{\eta,\theta}|\eta - \tilde{\eta}| > 0,$$

where the infimum is over all selectors $\tilde{\eta}$ of a vector η in model (16).

Remark 2. We should remark that the best selection procedure yields exact variable selection only if the condition $\liminf_{\varepsilon \to 0} \frac{u_\varepsilon(r_\varepsilon)}{\sqrt{2\log d} + \sqrt{2\log s}} > 1$ holds; at the same time, the best selection procedure gives almost full variable selection if a milder condition $\liminf_{\varepsilon \to 0} \frac{u_\varepsilon(r_\varepsilon)}{\sqrt{2\log(d/s)}} > 1$ is met.

5 Proofs of the Theorems

In this section, we prove Theorems 3 and 4. The proofs of Theorems 5 and 6 go along the same lines and therefore are omitted. Throughout the proof, the exponential bounds (26) and (27) on the tail probabilities of the statistics $t_j(s)$ will be frequently used.

Proof of Theorem 3. Let $m_0 \in \{2, \ldots, M\}$ be such that

$$s_{m_0-1} \le s < s_{m_0},$$

which implies that $s_{m_0}/s < d^{\Delta}$. Then, using the definition of the selector $\hat{\eta}(s_{\hat{m}})$, we can write

$$\begin{aligned}
&\sup_{\eta\in\mathcal{H}_{d,s}} \sup_{\theta\in\Theta_{\sigma,d}(r_\varepsilon)} s^{-1}\mathbf{E}_{\eta,\theta}|\hat{\eta}(s_{\hat{m}})-\eta| \\
&\le \sup_{\eta\in\mathcal{H}_{d,s}} \sup_{\theta\in\Theta_{\sigma,d}(r_\varepsilon)} s^{-1}\mathbf{E}_{\eta,\theta}\left(|\hat{\eta}(s_{\hat{m}})-\eta||\hat{m}<m_0\right)\mathbf{P}_{\eta,\theta}\left(\hat{m}<m_0\right) \\
&+ \sup_{\eta\in\mathcal{H}_{d,s}} \sup_{\theta\in\Theta_{\sigma,d}(r_\varepsilon)} s^{-1}\mathbf{E}_{\eta,\theta}\left(|\hat{\eta}(s_{\hat{m}})-\eta||\hat{m}\ge m_0\right)\mathbf{P}_{\eta,\theta}\left(\hat{m}\ge m_0\right) \\
&\le \sup_{\eta\in\mathcal{H}_{d,s}} \sup_{\theta\in\Theta_{\sigma,d}(r_\varepsilon)} s^{-1}\mathbf{E}_{\eta,\theta}\left(|\hat{\eta}(s_{\hat{m}})-\eta||\hat{m}<m_0\right)\mathbf{P}_{\eta,\theta}\left(\hat{m}<m_0\right) \\
&\qquad + \sup_{\eta\in\mathcal{H}_{d,s}} \sup_{\theta\in\Theta_{\sigma,d}(r_\varepsilon)} (d/s)\mathbf{P}_{\eta,\theta}\left(\hat{m}\ge m_0\right) =: I_1 + I_2. \qquad (38)
\end{aligned}$$

To complete the proof, we need to show that I_1 and I_2 are both negligibly small when ε is small.

Consider the term I_1 and observe that for all $\eta \in \mathcal{H}_{d,s}$ and $\theta \in \Theta_{\sigma,d}(r_\varepsilon)$,

$$\begin{aligned}
&s^{-1}\mathbf{E}_{\eta,\theta}\left(|\hat{\eta}(s_{\hat{m}})-\eta||\hat{m}<m_0\right)\mathbf{P}_{\eta,\theta}\left(\hat{m}<m_0\right) \\
&\le s^{-1}\mathbf{E}_{\eta,\theta}\left(|\hat{\eta}(s_{\hat{m}})-\hat{\eta}(s_{m_0})|\ \hat{m}<m_0\right) + s^{-1}\mathbf{E}_{\eta,\theta}\left(|\hat{\eta}(s_{m_0})-\eta||\hat{m}<m_0\right)\mathbf{P}_{\eta,\theta}\left(\hat{m}<m_0\right) \\
&\le s^{-1}v_{m_0} + s^{-1}\mathbf{E}_{\eta,\theta}|\hat{\eta}(s_{m_0})-\eta|,
\end{aligned}$$

where by (34) and the choice of the sequences τ_d and Δ

$$s^{-1}v_{m_0} = \tau_d^{-1}(s_{m_0}/s) < {\tau_d}^{-1}d^{\Delta} = o(1).$$

Next, by definition of the set $\mathcal{H}_{d,s}$ of s-sparse d-dimensional vectors η, we have

$$\begin{aligned}
\sup_{\eta\in\mathcal{H}_{d,s}} \sup_{\theta\in\Theta_{\sigma,d}(r_\varepsilon)} s^{-1}\mathbf{E}_{\eta,\theta}|\hat{\eta}(s_{m_0})-\eta| &\le (d/s)\mathbf{P}_0\left(t_1(s_{m_0}) > \sqrt{2\log(d/s_{m_0})+\delta\log d}\right) \\
&+ \sup_{\theta_1\in\Theta_\sigma(r_\varepsilon)} \mathbf{P}_{\theta_1}\left(t_1(s_{m_0}) \le \sqrt{2\log(d/s_{m_0})+\delta\log d}\right) \qquad (39)
\end{aligned}$$

where by (26) the first summand in the above expression satisfies

$$\begin{aligned}
&(d/s)\mathbf{P}_0\left(t_1(s_{m_0}) > \sqrt{2\log(d/s_{m_0})+\delta\log d}\right) \\
&\le (d/s)\exp\left(-\left(\log(d/s_{m_0})+(\delta/2)\log d\right)(1+o(1))\right) \\
&= O\left((s_{m_0}/s)d^{-\delta/2}\right) = O\left(d^{\Delta-\delta/2}\right) = o(1),
\end{aligned}$$

and the last equality is due to (21) and (34).

To treat the second term on the right side of (39), recall that $1 < s_{m_0}/s < d^{\Delta}$. Then, by the assumption on the parameter $r_\varepsilon = r_\varepsilon(s)$ and the 'continuity' of the function $u_\varepsilon(r_\varepsilon)$ as stated in (7), using the fact that $\Delta \log d \to 0$ as $d \to \infty$, one can find a constant $\delta_1 > 0$ such that for all sufficiently small ε

$$r_\varepsilon \geq r^*_\varepsilon(s_{m_0})(1+\delta_1).$$

From this, using Proposition 4.1 in [5] and recalling formula (24),

$$\begin{aligned}
\inf_{\theta_1 \in \Theta_\sigma(r_\varepsilon)} \mathbf{E}_{\theta_1}\left(t_1(s_{m_0})\right) &\geq \inf_{\theta_1 \in \Theta_\sigma(r^*_\varepsilon(s_{m_0})(1+\delta_1))} \mathbf{E}_{\theta_1}\left(t_1(s_{m_0})\right) \\
&\geq (1+\delta_1)^2 \inf_{\theta_1 \in \Theta_\sigma(r^*_\varepsilon(s_{m_0}))} \mathbf{E}_{\theta_1}\left(t_1(s_{m_0})\right) \geq (1+\delta_1)^2 u_\varepsilon(r^*_\varepsilon(s_{m_0})) \\
&= (1+\delta_1)^2 \sqrt{2\log(d/s_{m_0})} > \sqrt{2\log(d/s_{m_0}) + \delta \log d}, \qquad (40)
\end{aligned}$$

where the last inequality follows from the fact that $d^c \leq s_{m_0} < d^C$, which implies $\delta \log d = o(\log(d/s_{m_0}))$. Thus as $\varepsilon \to 0$

$$\sqrt{2\log(d/s_{m_0}) + \delta \log d} - \inf_{\theta_1 \in \Theta_\sigma(r_\varepsilon)} \mathbf{E}_{\theta_1}\left(t_1(s_{m_0})\right) \to -\infty. \qquad (41)$$

Now (27) in combination with (40) and (41) gives, uniformly in $\theta_1 \in \Theta_\sigma(r_\varepsilon)$,

$$\begin{aligned}
&\mathbf{P}_{\theta_1}\left(t_1(s_{m_0}) \leq \sqrt{2\log(d/s_{m_0}) + \delta \log d}\right) \\
&\leq \mathbf{P}_{\theta_1}\left(t_1(s_{m_0}) - \mathbf{E}_{\theta_1}\left(t_1(s_{m_0})\right) \leq \sqrt{2\log(d/s_{m_0}) + \delta \log d} - \inf_{\theta_1 \in \Theta_\sigma(r_\varepsilon)} \mathbf{E}_{\theta_1}\left(t_1(s_{m_0})\right)\right) \\
&\leq \mathbf{P}_{\theta_1}\left(t_1(s_{m_0}) - \mathbf{E}_{\theta_1}\left(t_1(s_{m_0})\right) \leq -\sqrt{2\log(d/s_{m_0})}\left[(1+\delta_1)^2 - 1 + o(1)\right]\right) \\
&\leq \exp\left(-\log(d/s_{m_0})\left[(1+\delta_1)^2 - 1 + o(1)\right]^2 (1+o(1))\right) \\
&= O\left((s_{m_0}/d)^{\left[(1+\delta_1)^2-1\right]^2}\right) = o(1).
\end{aligned}$$

Putting everything together, we conclude that the first term on the right side of (38) satisfies

$$I_1 = o(1), \quad \varepsilon \to 0. \qquad (42)$$

Let us now show that

$$I_2 = \sup_{\eta \in \mathcal{H}_{d,s}} \sup_{\theta \in \Theta_{\sigma,d}(r_\varepsilon)} (d/s)\mathbf{P}_{\eta,\theta}\left(\hat{m} \geq m_0\right) = o(1).$$

By definition of $\hat{m}$, for all $\eta \in \mathcal{H}_{d,s}$ and all $\theta \in \Theta_{\sigma,d}(r_\varepsilon)$,

$$\begin{aligned}
\mathbf{P}_{\eta,\theta}\left(\hat{m} \geq m_0\right) &= \sum_{k=m_0}^{M} \mathbf{P}_{\eta,\theta}\left(\hat{m} = k\right) \\
&= \sum_{k=m_0}^{M} \mathbf{P}_{\eta,\theta}\left(\exists\, i \in \{k, \ldots, M\} : |\hat{\eta}(s_{k-1}) - \hat{\eta}(s_i)| > v_i\right) \\
&\leq \sum_{k=m_0}^{M} \sum_{i=k}^{M} \mathbf{P}_{\eta,\theta}\left(|\hat{\eta}(s_{k-1}) - \hat{\eta}(s_i)| > v_i\right) \\
&= \sum_{k=m_0}^{M} \sum_{i=k}^{M} \mathbf{P}_{\eta,\theta}\left(\sum_{j=1}^{d} |\hat{\eta}_j(s_{k-1}) - \hat{\eta}_j(s_i)| > v_i\right).
\end{aligned}$$

Now, we introduce independent events

$$A_j(s) = \left\{ t_j(s) > \sqrt{2\log(d/s) + \delta \log d} \right\}, \quad j = 1, \ldots, d,$$

and denote by $\overline{A_j(s)}$ the complement of $A_j(s)$. Observing that for all $m_0 \le k \le i \le M$ the quantity $|\hat\eta_j(s_{k-1}) - \hat\eta_j(s_i)|$ is non-zero only if either $A_j(s_{k-1}) \cap \overline{A_j(s_i)}$ or $\overline{A_j(s_{k-1})} \cap A_j(s_i)$ occurs, we may continue

$$\mathbf{P}_{\eta,\theta}\left(\hat m \ge m_0\right) \le \sum_{k=m_0}^{M} \sum_{i=k}^{M} \mathbf{P}_{\eta,\theta} \left(\sum_{j=1}^{d} \left[\mathbb{I}\left(A_j(s_{k-1}) \cap \overline{A_j(s_i)} \right) + \mathbb{I}\left(\overline{A_j(s_{k-1})} \cap A_j(s_i) \right) \right] \right).$$

To bound this sum, we shall apply Bernstein's inequality saying that if $\mathbb{X}_1, \ldots, \mathbb{X}_d$ are independent random variables such that for all $j = 1, \ldots, d$ and for some $H > 0$

$$\mathbf{E}(\mathbb{X}_j) = 0 \quad \text{and} \quad \left|\mathbf{E}(\mathbb{X}_j^m)\right| \le \frac{\mathbf{E}(\mathbb{X}_j^2)}{2} H^{m-2} m! < \infty, \quad m = 2, 3, \ldots, \tag{43}$$

then (see, for example, pp. 164–165 of [2])

$$\max\left\{ \mathbf{P}\left(\mathbb{S}_d \ge t\right), \mathbf{P}\left(\mathbb{S}_d \le -t\right) \right\} \le \begin{cases} \exp\left(-t^2/4B_d^2\right) & \text{if} \quad 0 \le t \le B_d^2/H, \\ \exp\left(-t/4H\right) & \text{if} \quad t \ge B_d^2/H, \end{cases} \tag{44}$$

where $\mathbb{S}_d = \sum_{j=1}^d \mathbb{X}_j$ and $B_d^2 = \sum_{j=1}^d \mathbf{E}(\mathbb{X}_j^2)$. Observe that for independent random variables $\mathbb{X}_1, \ldots, \mathbb{X}_d$ with the property

$$\mathbf{E}(\mathbb{X}_j) = 0 \quad \text{and} \quad |\mathbb{X}_j| \le M, \quad j = 1, \ldots, d,$$

for some $M > 0$, the Bernstein condition (43) holds with $H = M/3$. Below we will use Bernstein's inequality in the case of $t \ge B_d^2/H$.

To do this, let us introduce random variables $\mathbb{X}_j = \mathbb{X}_j(s_{k-1}, s_i)$, $1 \le j \le d$, $m_0 \le k \le M$, $k \le i \le M$, by the formula

$$\begin{aligned} \mathbb{X}_j \;\; = \;\; & \mathbb{I}\left(A_j(s_{k-1}) \cap \overline{A_j(s_i)} \right) + \mathbb{I}\left(\overline{A_j(s_{k-1})} \cap A_j(s_i) \right) \\ & - \left[\mathbf{P}_{\eta,\theta}\left(A_j(s_{k-1}) \cap \overline{A_j(s_i)} \right) + \mathbf{P}_{\eta,\theta}\left(\overline{A_j(s_{k-1})} \cap A_j(s_i) \right) \right], \end{aligned}$$

and observe that $|\mathbb{X}_j| \le 4$, $j = 1, \ldots, d$, and for all $\eta \in \mathcal{H}_{d,s}$ and $\theta \in \Theta_{\sigma,d}(r_\varepsilon)$

$$\mathbf{E}_{\eta,\theta}(\mathbb{X}_j) = 0, \quad j = 1, \ldots, d.$$

Before applying Bernstein's inequality, we show that for all $\eta \in \mathcal{H}_{d,s}$ and $\theta \in \Theta_{\sigma,d}(r_\varepsilon)$, and for all $m_0 \le k \le M$ and $k \le i \le M$

$$\sum_{j=1}^{d} \left[\mathbf{P}_{\eta,\theta}\left(A_j(s_{k-1}) \cap \overline{A_j(s_i)} \right) + \mathbf{P}_{\eta,\theta}\left(\overline{A_j(s_{k-1})} \cap A_j(s_i) \right) \right] = o(v_i). \tag{45}$$

We have

$$\sup_{\eta\in\mathcal{H}_{d,s}} \sup_{\theta\in\Theta_{\sigma,d}(r_\varepsilon)} \sum_{j=1}^{d} \left(\mathbf{P}_{\eta,\theta}\left(A_j(s_{k-1})\cap\overline{A_j(s_i)}\right)+\mathbf{P}_{\eta,\theta}\left(\overline{A_j(s_{k-1})}\cap A_j(s_i)\right)\right)$$
$$=(d-s)\left[\mathbf{P}_0\left(A_1(s_{k-1})\cap\overline{A_1(s_i)}\right)+\mathbf{P}_0\left(\overline{A_1(s_{k-1})}\cap A_1(s_i)\right)\right]$$
$$+s\sup_{\theta_1\in\Theta_\sigma(r_\varepsilon)}\left[\mathbf{P}_{\theta_1}\left(A_1(s_{k-1})\cap\overline{A_1(s_i)}\right)+\mathbf{P}_{\theta_1}\left(\overline{A_1(s_{k-1})}\cap A_1(s_i)\right)\right]$$
$$\le d\left[\mathbf{P}_0\left(t_1(s_{k-1})>\sqrt{2\log(d/s_{k-1})+\delta\log d}\right)+\mathbf{P}_0\left(t_1(s_i)>\sqrt{2\log(d/s_i)+\delta\log d}\right)\right]$$
$$+s\sup_{\theta_1\in\Theta_\sigma(r_\varepsilon)}\left[\mathbf{P}_{\theta_1}\left(t_1(s_{k-1})\le\sqrt{2\log(d/s_{k-1})+\delta\log d}\right)+\mathbf{P}_{\theta_1}\left(t_1(s_i)\le\sqrt{2\log(d/s_i)+\delta\log d}\right)\right]$$
$$=: J_1(s_{k-1},s_i)+J_2(s_{k-1},s_i). \tag{46}$$

Recalling (26) and the relation $\tau_d d^{-\delta/2}\to 0$ as $d\to\infty$, we have

$$d\mathbf{P}_0\left(t_1(s_i)>\sqrt{2\log(d/s_i)+\delta\log d}\right)\le d\exp\left(-\left(\log(d/s_i)+(\delta/2)\log d\right)(1+o(1))\right)$$
$$=O\left(s_i d^{-\delta/2}\right)=O\left(v_i\tau_d d^{-\delta/2}\right)=o(v_i).$$

Similarly, using the fact that $v_{k-1}<v_i$ when $k\le i\le M$, we obtain

$$d\mathbf{P}_0\left(t_1(s_{k-1})>\sqrt{2\log(d/s_{k-1})+\delta\log d}\right)=o(v_{k-1})=o(v_i).$$

Therefore for all $m_0\le k\le M$ and $k\le i\le M$

$$J_1(s_{k-1},s_i)=o(v_i). \tag{47}$$

Consider the second term on the right side of (46), $J_2(s_{k-1},s_i)$. First, note that for all $m_0\le k\le M$ and $k\le i\le M$,

$$s<s_i \quad\text{and}\quad s<s_{k-1},\quad k\neq m_0.$$

and for $k=m_0$ one has $s_{k-1}=s_{m_0-1}\le s$, which implies $s/s_{m_0-1}<d^{\Delta}$. Therefore, by the assumption on $r_\varepsilon=r_\varepsilon(s)$ and the 'continuity' of the function $u_\varepsilon(r_\varepsilon)$ as cited in (7), using the fact that $\Delta\log d\to 0$ as $d\to\infty$, one can find constants $\delta_2>0$ and $\delta_3>0$ such that for all sufficiently small ε

$$r_\varepsilon\ge r^*_\varepsilon(s_i)(1+\delta_2)\quad\text{and}\quad r_\varepsilon\ge r^*_\varepsilon(s_{k-1})(1+\delta_3)$$

when $m_0\le k\le M$ and $k\le i\le M$. From this, for all sufficiently small ε, cf. (40),

$$\inf_{\theta_1\in\Theta_\sigma(r_\varepsilon)}\mathbf{E}_{\theta_1}\left(t_1(s_i)\right)\ge(1+\delta_2)^2\sqrt{2\log(d/s_i)}>\sqrt{2\log(d/s_i)+\delta\log d}, \tag{48}$$

and hence as $\varepsilon\to 0$

$$\sqrt{2\log(d/s_i)+\delta\log d}-\inf_{\theta_1\in\Theta_\sigma(r_\varepsilon)}\mathbf{E}_{\theta_1}\left(t_1(s_i)\right)\to-\infty. \tag{49}$$

It now follows from (27), (48), and (49) that, uniformly in $\theta_1 \in \Theta_\sigma(r_\varepsilon)$,

$$
\begin{aligned}
& s\mathbf{P}_{\theta_1}\left(t_1(s_i) \leq \sqrt{2\log(d/s_i) + \delta \log d}\right) \\
\leq\ & s\mathbf{P}_{\theta_1}\left(t_1(s_i) - \mathbf{E}_{\theta_1}\left(t_1(s_i)\right) \leq \sqrt{2\log(d/s_i) + \delta \log d} - \inf_{\theta_1 \in \Theta_\sigma(r_\varepsilon)} \mathbf{E}_{\theta_1}\left(t_1(s_i)\right)\right) \\
\leq\ & s\mathbf{P}_{\theta_1}\left(t_1(s_i) - \mathbf{E}_{\theta_1}\left(t_1(s_i)\right) \leq -\sqrt{2\log(d/s_i)}\left[(1+\delta_2)^2 - 1 + o(1)\right]\right) \\
\leq\ & s\exp\left(-\log(d/s_i)\left[(1+\delta_2)^2 - 1 + o(1)\right]^2(1+o(1))\right) \\
=\ & O\left(s(s_i/d)^{[(1+\delta_2)^2-1]^2}\right) = O\left(s_i(s_i/d)^{[(1+\delta_2)-1]^2}\right) = o(v_i).
\end{aligned}
$$

Also, as relation (49) continues to hold with s_{k-1}, $m_0 \leq k \leq M$, instead of s_i, similar arguments yield

$$
s\mathbf{P}_{\theta_1}\left(t_1(s_{k-1}) \leq \sqrt{2\log(d/s_{k-1}) + \delta \log d}\right) = o(v_{k-1}) = o(v_i),
$$

which implies

$$
J_2(s_{k-1}, s_i) = o(v_i). \tag{50}
$$

Combining (46), (47) and (50), we arrive at (45). We see then by (45) that

$$
\begin{aligned}
\sum_{j=1}^{d} \mathbf{E}_{\eta,\theta}(\mathbb{X}_j^2) &= \left(\sum_{j=1}^{d}\left[\mathbf{P}_{\eta,\theta}\left(A_j(s_{k-1}) \cap \overline{A_j(s_i)}\right) + \mathbf{P}_{\eta,\theta}\left(\overline{A_j(s_{k-1})} \cap A_j(s_i)\right)\right]\right)(1+o(1)) \\
&= o(v_i).
\end{aligned}
$$

Therefore, the use of Bernstein's inequality as in (44) for the case of $t \geq B_d^2/H$ with $H = 4/3$ gives

$$
\begin{aligned}
I_2 &= \sup_{\eta \in \mathcal{H}_{d,s}} \sup_{\theta \in \Theta_{\sigma,d}(r_\varepsilon)} (d/s)\mathbf{P}_{\eta,\theta}\left(\hat{m} \geq m_0\right) \\
&\leq \sup_{\eta \in \mathcal{H}_{d,s}} \sup_{\theta \in \Theta_{\sigma,d}(r_\varepsilon)} (d/s) \sum_{k=m_0}^{M} \sum_{i=k}^{M} \mathbf{P}_{\eta,\theta}\left(\sum_{j=1}^{d} \mathbb{X}_j > v_i(1+o(1))\right) \\
&\leq (d/s) \sum_{k=m_0}^{M} \sum_{i=k}^{M} \exp\left(-(3v_i/16)(1+o(1))\right) = O\left(M^2(d/s)\exp\left(-(3/16)v_{m_0}\right)\right) \\
&= O\left(M^2(d/s)\exp\left(-(3d^c/16\tau_d)\right)\right) = o(1).
\end{aligned}
$$

This in combination with (38) and (42) completes the proof of Theorem 3. □

5.1 Proof of Theorem 4

To prove the theorem, we first pick good prior distributions on $\eta = (\eta_j)$ and $\theta = (\theta_j)$. Having done this, we bound the normalized minimax risk by the normalized Bayes risk and show that the latter is strictly positive. The first part of the proof up to relation (55) go along the lines of that of Theorem 2 in [13], with $p = s/d$ instead of $p = d^{-\beta}$.

Let $\theta_j^* = (\theta_{j,k}^*)_{k\in\mathbb{Z}}$ be the extremal sequence in the problem (the same for all $j = 1, \dots, d$)

$$\frac{1}{2\varepsilon^4}\sum_{k\in\mathbb{Z}}\theta_{j,k}^4 \to \inf_{\theta_j\in\Theta_\sigma(r_\varepsilon)}.$$

Let the prior distribution of a 'vector' $\theta = (\theta_1, \dots, \theta_d) \in \Theta_{\sigma,d}(r_\varepsilon)$ be of the form

$$\pi_\theta(d\theta) = \prod_{j=1}^d \pi_{\theta_j}(d\theta_j), \quad \pi_{\theta_j}(d\theta_j) = \prod_{1\le|k|\le K_\varepsilon}\left(\frac{\delta_{-\theta_{j,k}^*} + \delta_{\theta_{j,k}^*}}{2}\right)(d\theta_{j,k}),$$

where δ_x is the δ-measure that puts a pointmass 1 at x. Denote by

$$p = s/d$$

the portion of non-zero components of vector $\eta = (\eta_1, \dots, \eta_d) \in \mathcal{H}_{d,s}$. The prior distribution of η is naturally defined to be

$$\pi_\eta(d\eta) = \prod_{j=1}^d \pi_{\eta_j}(d\eta_j), \quad \pi_{\eta_j}(d\eta_j) = ((1-p)\delta_0 + p\delta_1)(d\eta_j).$$

Then, assuming that $\theta = (\theta_j)$ and $\eta = (\eta_j)$ are independent, we get

$$\begin{aligned} R_\varepsilon &:= \inf_{\tilde\eta}\sup_{\eta\in\mathcal{H}_{d,s}}\sup_{\theta\in\Theta_{\sigma,d}(r_\varepsilon)} s^{-1}\mathbf{E}_{\eta,\theta}|\eta-\tilde\eta| \ge s^{-1}\inf_{\tilde\eta}\mathbf{E}_{\pi_\eta}\mathbf{E}_{\pi_\theta}\mathbf{E}_{\eta,\theta}|\eta-\tilde\eta| \\ &= s^{-1}\inf_{\tilde\eta}\mathbf{E}_{\pi_\eta}\mathbf{E}_{\pi_\theta}\mathbf{E}_{\eta,\theta}\sum_{j=1}^d|\eta_j-\tilde\eta_j| = s^{-1}\inf_{\tilde\eta}\sum_{j=1}^d\mathbf{E}_{\pi_{\eta_j}}\mathbf{E}_{\pi_{\theta_j}}\mathbf{E}_{\eta_j\theta_j}|\eta_j-\tilde\eta_j|, \end{aligned}$$

where the infimum is over all selectors $\tilde\eta = (\tilde\eta_j)$ and $\mathbf{E}_{\eta_j\theta_j}$ is the expected value that corresponds to the measure $\mathbf{P}_{\eta_j\theta_j}$ induced by the observation $X_j = (X_{j,k})_{1\le|k|\le K_\varepsilon}$ consisting of independent random variables $X_{j,k}$ that follow normal distributions $N(\eta_j\theta_{j,k}, \varepsilon^2)$.

Consider the mixture of distributions given by the formula

$$\mathbf{P}_{\pi,\eta_j}(dX_j) = \mathbf{E}_{\pi_{\theta_j}}\mathbf{P}_{\eta_j\theta_j}(dX_{j,k}) = \prod_{1\le|k|\le K_\varepsilon}\left(\frac{N(-\eta_j\theta_{j,k}^*, \varepsilon^2) + N(\eta_j\theta_{j,k}^*, \varepsilon^2)}{2}\right)(dX_{j,k}). \quad (51)$$

In particular, when $\eta_j = 0$, $\mathbf{P}_{\pi,0}(dX_j) = \prod_{1\le|k|\le K_\varepsilon} N(0,\varepsilon^2)(dX_{j,k})$. Using the notation

$$v_{j,k}^* = \frac{\theta_{j,k}^*}{\varepsilon}, \quad (52)$$

we obtain with respect to the probability measure $\mathbf{P}_{\pi,\eta_j}$

$$Y_{j,k} := \frac{X_{j,k}}{\varepsilon} = \eta_j v_{j,k}^* + \xi_{j,k} \overset{\text{ind.}}{\sim} N(\eta_j v_{j,k}^*, 1), \quad 1\le j\le d, \quad 1\le|k|\le K_\varepsilon.$$

Next, denoting $Y_j = (Y_{j,k})_{1\le|k|\le K_\varepsilon}$, we may rewrite the likelihood ratio in the form

$$\frac{d\mathbf{P}_{\pi,\eta_j}}{d\mathbf{P}_{\pi,0}}(Y_j) = \prod_{1\le|k|\le K_\varepsilon}\exp\left(-\frac{\eta_j(v_{j,k}^*)^2}{2}\right)\cosh\left(\eta_j v_{j,k}^* Y_{j,k}\right). \quad (53)$$

From this, using the fact that each η_j takes on only two values, zero and one, with respective probabilities $(1-p)$ and p, we may continue

$$R_\varepsilon \;\geq\; s^{-1}\sum_{j=1}^{d}\inf_{\tilde\eta_j}\mathbf{E}_{\pi_{\eta_j}}\mathbf{E}_{\pi,\eta_j}|\eta_j \,-\, \tilde\eta_j| \;=\; s^{-1}\sum_{j=1}^{d}\inf_{\tilde\eta_j}\left[(1-p)\mathbf{E}_{\pi,0}(\tilde\eta_j)+p\mathbf{E}_{\pi,1}(1-\tilde\eta_j)\right], \quad (54)$$

where $\inf_{\tilde\eta_j}(1-p)\mathbf{E}_{\pi,0}(\tilde\eta_j)+p\mathbf{E}_{\pi,1}(1-\tilde\eta_j)$ is the Bayes risk in the problem of testing two simple hypotheses

$$H_0 : \mathbf{P} = \mathbf{P}_{\pi,0} \quad \text{vs.} \quad H_1 : \mathbf{P} = \mathbf{P}_{\pi,1},$$

with the probability measures $\mathbf{P}_{\pi,0}$ and $\mathbf{P}_{\pi,1}$ defined according to (51). In particular, under the null hypothesis, the vector $Y_j = (Y_{j,k})_{1\le|k|\le K_\varepsilon}$ has a normal distribution with density function $p_{\pi,0}(t) = \prod\limits_{1\le|k|\le K_\varepsilon}(2\pi)^{-1/2}\exp(-t_k^2/2)$, $t=(t_k)_{1\le|k|\le K_\varepsilon}$. By (53) the likelihood ratio in this problem becomes

$$\Lambda_\pi = \Lambda_\pi(Y_j) = \frac{d\mathbf{P}_{\pi,1}}{d\mathbf{P}_{\pi,0}}(Y_j) = \prod_{1\le|k|\le K_\varepsilon}\exp\left(-\frac{(v^*_{j,k})^2}{2}\right)\cosh\left(v^*_{j,k}Y_{j,k}\right),$$

and the optimal (Bayes) test η_B that minimizes the Bayes risk in hand has the form (see, for example, [4, Sec. 8.11])

$$\eta_B(Y_j) = \mathbb{I}\left(\Lambda_\pi(Y_j) \ge \frac{1-p}{p}\right).$$

Using this, we infer from (54) that

$$\begin{aligned} R_\varepsilon &= \inf_{\tilde\eta}\sup_{\eta\in\mathcal{H}_{d,s}}\sup_{\theta\in\Theta_{\sigma,d}(r_\varepsilon)} s^{-1}\mathbf{E}_{\eta,\theta}|\eta-\tilde\eta| \\ &\ge (d/s)\mathbf{P}_{\pi,0}\left(\Lambda_\pi(Y_1)\ge\frac{1-p}{p}\right)+\mathbf{P}_{\pi,1}\left(\Lambda_\pi(Y_1)<\frac{1-p}{p}\right) =: A_\varepsilon + B_\varepsilon. \qquad (55)\end{aligned}$$

where, under the $\mathbf{P}_{\pi,\eta_1}$-probability with $\eta_1\in\{0,1\}$, the vector $Y_1=(Y_{1,k})_{1\le|k|\le K_\varepsilon}$ has independent normal components

$$Y_{1,k} = \eta_1 v^*_{1,k} + \xi_{1,k} \sim N(\eta_1 v^*_{1,k},1), \quad 1\le|k|\le K_\varepsilon.$$

It now follows from (55) that the minimax risk R_ε is positive if at least one of the terms, A_ε or B_ε, is positive. Let us prove that for all sufficiently small ε the probability B_ε is separated from zero.

Recall that $d=d_\varepsilon\to\infty$ and $s=s_d=o(d)$ as $\varepsilon\to0$. Put

$$H = H_\varepsilon = \log\left(\frac{1-p}{p}\right) \sim \log(d/s),$$

and introduce the random variable

$$\lambda_\pi = \lambda_\pi(Y_1) := \log\Lambda_\pi(Y_1).$$

Using the notation $\mathbf{P}_0$ for $\mathbf{P}_{\pi,0}$, consider the probability measure $\mathbf{P}_h$, depending on a positive parameter $h=h_\varepsilon$, that is defined by the formula

$$\frac{d\mathbf{P}_h}{d\mathbf{P}_0}(Y_1) := \frac{\exp(h\lambda_\pi(Y_1))}{\Psi(h)}, \quad \Psi(h) = \mathbf{E}_{\mathbf{P}_0}\exp(h\lambda_\pi(Y_1)).$$

With the parameter $h > 0$ chosen to satisfy

$$\mathbf{E}_{\mathbf{P}_h}\lambda_\pi = H,$$

we have (see Lemma 2 in [13])

$$h \sim \frac{1}{2} + \frac{H}{u_\varepsilon^2} = O(1), \tag{56}$$

and (see formula (45) in [13])

$$\Psi(h) = \exp\left(\frac{h^2 - h}{2}\, u_\varepsilon^2(1 + o(1))\right), \tag{57}$$

where for notational simplicity we use u_ε^2 for $u_\varepsilon^2(r_\varepsilon)$.

We have

$$\begin{aligned} B_\varepsilon &= \mathbf{E}_{\pi,1}\left(\mathbb{I}\left(\lambda_\pi(Y_1) < H\right)\right) = \mathbf{E}_{\pi,0}\left(\exp(\lambda_\pi(Y_1))\mathbb{I}\left(\lambda_\pi(Y_1)\right) < H\right)\right) \\ &= \mathbf{E}_h\left(\frac{d\mathbf{P}_0}{d\mathbf{P}_h}(Y_1)\exp(\lambda_\pi(Y_1))\mathbb{I}\left(\lambda_\pi(Y_1)\right) < H\right)\right) \\ &= \Psi(h)\mathbf{E}_h\left(\exp[(1-h)\lambda_\pi(Y_1)]\mathbb{I}\left(\lambda_\pi(Y_1)\right) < H\right)\right). \end{aligned} \tag{58}$$

By Lemma 3 in [13], the standardized random variable

$$Z_h := \frac{\lambda_\pi - \mu_h}{\sigma_h},$$

where

$$\mu_h = \mathbf{E}_{\mathbf{P}_h}(\lambda_\pi) = u_\varepsilon^2(h - 1/2)(1 + o(1)), \quad \sigma_h^2 = \mathbf{Var}_{\mathbf{P}_h}(\lambda_\pi) = u_\varepsilon^2(1 + o(1)),$$

converges in $\mathbf{P}_h$-distribution to an $N(0,1)$. Therefore the statistic $\lambda_\pi(Y_1)$ on the right side of (58) is nearly a normal $N(H, u_\varepsilon^2)$ random variable.

Next, by assumption and the 'continuity' of u_ε as stated in (7), for some constant $\delta_1 > 0$

$$u_\varepsilon/\sqrt{\log(d/s)} \le \sqrt{2}(1 - \delta_1),$$

provided ε is small enough. This and formula (56) give the inequality $1 - h < 0$, which implies for all $y \in \mathbb{R}^{2K_\varepsilon}$ and all sufficiently small ε

$$\exp[(1-h)\lambda_\pi(y)]\mathbb{I}\left(\lambda_\pi(y)\right) < H) \le \exp\left[(1-h)H\right] \sim (d/s)^{1-h} \le \text{const}.$$

Then, by the dominant convergence theorem, the replacement of $\lambda_\pi(Y_1)$ by an $N(H, u_\varepsilon^2)$ on the right side (58) and the use of (56) and (57) yield for all sufficiently small ε

$$\begin{aligned} B_\varepsilon &\sim \exp\left(\frac{h^2-h}{2}u_\varepsilon^2\right)\int_{-\infty}^{H} \exp\left[(1-h)x\right]\frac{1}{\sqrt{2\pi}u_\varepsilon}\exp\left(-\frac{(x-H)^2}{2u_\varepsilon^2}\right)dx \\ = \exp\left(\frac{h^2-h}{2}u_\varepsilon^2 + H(1-h) + \frac{(1-h)^2u_\varepsilon^2}{2}\right)&\int_{-\infty}^{H}\frac{1}{\sqrt{2\pi}u_\varepsilon}\exp\left(-\frac{\left(x-(H+(1-h)u_\varepsilon^2)\right)^2}{2u_\varepsilon^2}\right)dx \\ &\sim \exp(0)\int_{-\infty}^{H}\frac{1}{\sqrt{2\pi}u_\varepsilon}\exp\left(-\frac{\left(x-(H+(1-h)u_\varepsilon^2)\right)^2}{2u_\varepsilon^2}\right)dx \\ &\ge \int_{-\infty}^{H+(1-h)u_\varepsilon^2}\frac{1}{\sqrt{2\pi}u_\varepsilon}\exp\left(-\frac{\left(x-(H+(1-h)u_\varepsilon^2)\right)^2}{2u_\varepsilon^2}\right)dx = 1/2. \end{aligned}$$

From this

$$\liminf_{\varepsilon\to 0} R_\varepsilon \geq \liminf_{\varepsilon\to 0} B_\varepsilon \geq 1/2 > 0,$$

and the proof of Theorem 4 is complete. □

6 Concluding remarks

In the context of variable selection in high dimensions, in both regression and white noise settings, simple thresholding provides plausible alternative to the lasso for a large range of problems. As a statistical tool, thresholding strategy is simple in nature and is not as computationally demanding as the lasso, especially in very high dimensional problems. At the same time, it is capable of doing at least as good as the lasso, or even better (see our Theorems 1 to 6, Theorems 9 to 11 in [6], and Theorems 1 and 2 in [13] for details). In light of these facts, we support the viewpoint of Genovese et al. [6] that for sparse high-dimensional regression problems a simple thresholding procedure merits further investigation.

To conclude our study, we point out possible directions for extending the results obtained in this paper. For the two function spaces $\mathcal{F}_\sigma$ at hand, it might be of interest to produce asymptotically exact and almost full selectors in very high dimensional settings when the conditions $\log d = o\left(\varepsilon^{-2/(2\sigma+1)}\right)$ and $\log d = o(\log \varepsilon^{-1})$ on the growth of d as a function of ε are violated.

The setup of inverse problems, where the observations are $X_\varepsilon = Kf + \varepsilon W$, with K being a linear operator such that $K^\star K$ is compact, translates into a Gaussian sequence model with heterogenous observations $X_{j,k} = \eta_j \theta_{j,k} + \epsilon v_k \xi_{j,k}$, where v_k^{-2} are the eigenvalues of $K^\star K$. This case, which extends our setup, can be treated by using the sharp testing results for the inverse problems obtained in [14].

Furthermore, handling the problem of variable selection in a sequence space model, general ellipsoids $\{\theta \in l_2(\mathbb{Z}) : \sum_{k\in\mathbb{Z}} c_k^2 \theta_k^2 \leq 1\}$ in $l_2(\mathbb{Z})$, with semi-axes c_k decreasing fast enough, could be studied. A more complicated model, in which a d-variate regression function f admits a decomposition to a sum of k-variate components, with $k \geq 2$ and only a small number s of these components being non-zero, also deserves some attention.

Eliminating the assumption of *known* parameter σ leads to the problem of adapting the proposed selection procedures to the possible values of σ.

To pursue more practical goals, one can try to translate the results obtained for an additive s-sparse Gaussian white noise model to the corresponding discrete regression model for which the corresponding detection problem was solved in [1].

References

[1] F. Abramovich, I. De Feis, T. Sapatinas, *Optimal testing for additivity in multiple nonparametric regression.* — Annals of the Insitute of Statistical Mathematics **61**, No. 3 (2009), 691–714.

[2] S. N. Bernstein, *Probability Theory*, OGIZ, Moscow–Leningrad (1946). In Russian.

[3] L. Comminges and A. S. Dalalyan, *Tight conditions for consistency of variable selection in the context of high dimensionality.* — Annals of Statistics **40**, No. 5 (2012), 2667–2696.

[4] M. De Groot, *Optimal Statistical Decisions.* McGraw-Hill Book Company, New York (1970).

[5] G. Gayraud and Yu. I. Ingster, *Detection of sparse variable functions.* — Electronic Journal of Statistics **6** (2012), 1409–1448.

[6] C. R. Genovese, J. Jin, L. Wasserman, and Z. Yao, *A comparison of the lasso and marginal regression.* — Journal of Machine Learning Research **13** (2012), 2107–2143.

[7] Y. K. Golubev and B. Y. Levit, *Asymptotically efficient estimation for analytic distributions.* — Mathematical Methods of Statisics **3** (1996), 357–368.

[8] J. Huang, J. L. Horowitz, and F. Wei, *Variable selection in nonparametric additive models.* — Annals of Statistics **38** (2010), 2282–2313.

[9] I. A. Ibragimov and R. Z. Khasminskii, *Some estimation problems on infinite dimensional Gaussian white noise.* In Festschrift for Lucien Le Cam. Research Papers in Probability and Statistics, Springer-Verlag, New York (1997), 275–296.

[10] Yu. I. Ingster, *Asymptotically minimax hypothesis testing for nonparametric alternatives. I.* — Mathematical Methods of Statistics **2**, No. 2 (1993), 85–114.

[11] Yu. I. Ingster, *Asymptotically minimax hypothesis testing for nonparametric alternatives. II.* — Mathematical Methods of Statistics **2**, No. 3 (1993), 171–189.

[12] Yu. I. Ingster, *Asymptotically minimax hypothesis testing for nonparametric alternatives. III.* — Mathematical Methods of Statistics **2**, No. 4 (1993), 249–268.

[13] Yu. I. Ingster and N. A. Stepanova, *Adaptive variable selection in nonparametric sparse regression.* — Journal of Mathematical Sciences **199**, No. 2 (2014), 184–201.

[14] Yu. I. Ingster, T. Sapatinas, and I. A. Suslina, *Minimax signal detection in ill-posed inverse problems.* — Annals of Statistics **40**, No. 3 (2012), 1524–1549.

[15] Yu. I. Ingster and I. A. Suslina, *Nonparametric Goodness-of-Fit Testing Under Gaussian Models.* Lect. Notes Statist., Vol. **169**, Springer-Verlag, New York (2003).

[16] Yu. I. Ingster and I. A. Suslina, *On estimation and detection of smooth function of many variables.* — Mathematical Methods of Statistics **14** (2005), 299–331.

[17] O. V. Lepski, *One problem of adaptive estimation in Gaussian white noise.* — Theory of Probababaility and Its Applications **35** (1990), 459–470.

[18] G. Raskutti, M. J. Wainwright, and B. Yu, *Minimax-optimal rates for high-dimensional sparse additive models over kernel classes.* — Journal of Machine Learning Research **13** (2012), 281–319.

[19] A. V. Skorohod, *Integration in Hilbert Spaces*, Springer-Verlag, Berlin–New York (1974).